Int J of Sci and Math Educ (2017) 15:663–682
DOI 10.1007/s10763-016-9723-0

Measuring Critical Thinking in Physics: Development and Validation of a Critical Thinking Test in Electricity and Magnetism

Dawit Tibebu Tiruneh[1] · **Mieke De Cock**[2] · **Ataklti G. Weldeslassie**[3] · **Jan Elen**[1] · **Rianne Janssen**[4]

Received: 16 December 2015 / Accepted: 14 February 2016 / Published online: 29 February 2016

Abstract Although the development of critical thinking (CT) is a major goal of science education, adequate emphasis has not been given to the measurement of CT skills in specific science domains such as physics. Recognizing that adequately assessing CT implies the assessment of both domain-specific and domain-general CT skills, this study reports on the development and validation of a test designed to measure students' acquisition of CT skills in electricity and magnetism (CTEM). The CTEM items were designed to mirror the structural components of items identified in an existing standardized domain-general CT test, and targeted content from an introductory Electricity and Magnetism (E&M) course. A preliminary version of the CTEM test was initially piloted on three groups of samples: interviews with physics experts ($N=3$), student cognitive interviews ($N=6$), and small-scale paper and pencil administration ($N=19$). Modifications were made afterwards and the test was administered to a different group of second-year students whose major was mechanical engineering ($N=45$). The results showed that the internal consistency (Cronbach's $\alpha=.72$) and inter-rater reliability (Cohen's kappa = .83) of the CTEM test are acceptable. The findings overall suggest that the CTEM test can be used to measure the acquisition of domain-specific CT skills in E&M, and a good basis for future empirical research that focuses on the integration of CT skills within specific subject matter instruction. A broader CT assessment

✉ Dawit Tibebu Tiruneh
dawittibebu.tiruneh@ppw.kuleuven.be

[1] Centre for Instructional Psychology and Technology, KU Leuven, Dekenstraat 2, Box 3773, 3000 Leuven, Belgium

[2] Department of Physics and Astronomy & LESEC, KU Leuven, Leuven, Belgium

[3] Science, Engineering and Technology Group, KU Leuven Group T, Leuven, Belgium

[4] Educational Effectiveness and Evaluation, KU Leuven, Leuven, Belgium

framework is proposed and possible research questions that can be addressed through the CTEM test are discussed.

Keywords Assessment · Domain-specific critical thinking · Physics · Science education

The development of critical thinking (CT) is widely claimed as a primary goal of science education (Adey & Shayer, 1994; Bailin, 2002; Siegel, 1988). CT involves the ability to draw valid inferences, identify relationships, analyze probabilities, make predictions and logical decisions, and solve complex problems (Halpern, 2014; Pascarella & Terenzini, 2005). Proficiency in CT is associated with success in undergraduate education, improved decision-making with regard to complex real-life problems, and more generally with a tendency to become a more active and informed citizen (Halpern, 2014). Accordingly, comprehensive science curricula revisions that focus on student acquisition of CT skills have long been called for by several stakeholders in education (Association of American Colleges and Universities [AAC&U], 2005; Facione, 1990a; Kuhn, 1999).

Most previous efforts to address the challenge of CT development took place in a context in which general CT skills were taught separately from regular subject matter domains (Ennis, 1989; Pascarella & Terenzini, 2005). However, this approach has become less dominant in recent years, and empirical attempts to develop CT have shifted mainly toward embedding CT skills within subject matter instruction (for reviews, see Niu, BeharHorenstein, & Garvan, 2013; Tiruneh, Verburgh, & Elen, 2014). The accompanying expectation has been that embedding CT skills within a subject matter instruction in various specific domains will facilitate the acquisition of CT skills that are applicable to a wide variety of thinking tasks within the domain in question and that it will facilitate their transfer to other problems in everyday life (Adey & Shayer, 1994; Lawson, 2004). Successful teaching of CT skills in coherence with the teaching of domain-specific knowledge is in other words expected to result in the development of both domain-specific and domain-general CT skills that are necessary to perform thinking tasks requiring a considerable mental activity such as predicting, analyzing, synthesizing, evaluating, reasoning, etc.

As the emphasis on the development of CT continues, the need for its assessment has become more crucial (Facione, 1990a; Halpern, 2010; Lin, 2014; Pascarella & Terenzini, 2005). At the same time, however, the notion of CT has been highly disputed, which in turn has led to confusion in its assessment. The lack of coherent and defensible conception of CT has been one of the major issues. The disagreement among educators and researchers with regard to the definition of CT and what is to be accomplished in assessing it effectively is widespread (Bailin, 2002; Ennis, 1993; Facione, 1990a). For example, Ennis (1993) defines CT as reasonable reflective thinking focused on deciding what to believe or do. Ennis (1993) argues that CT involves mental processes such as the ability to draw conclusions, judge the credibility of sources, develop and defend a position on an issue, ask appropriate clarifying questions, and plan experiments systematically. Halpern (2014) defines CT as the use of thinking strategies that increase the probability of a desirable outcome. Together with her definition, Halpern identified five major categories of CT skills that she argues as relevant for a desirable outcome in any domain: reasoning, hypothesis testing, argument analysis, likelihood and uncertainty analysis, and decision-making and problem-solving (Halpern, 2010, 2014).

There are diverse views held among scholars in terms of the core processes involved in CT. In an effort to assess CT proficiency, many tests were developed and validated. Ennis, Millman and Tomko (1985), for example, co-authored a domain-general CT test named the Cornell Critical Thinking Test (CCTT) that targets the following elements of CT: induction, deduction, credibility, prediction and experimental planning, and fallacies and assumption identification. On the other hand, Halpern, in line with her analysis of recent conceptions of CT, developed and validated a domain-general CT test named the Halpern Critical Thinking Assessment (HCTA; Halpern, 2010). The HCTA test focuses on measuring the five major elements of CT skills as identified by Halpern (2010): reasoning, argument analysis, hypothesis testing, likelihood and uncertainty analysis, and problem-solving and decision-making. It is possible to mention the focus of all the other existing domain-general CT tests, and one can reasonably conclude that all the domain-general CT tests are diverse in terms of their formats, scope, and psychometric characteristics. Such variations in CT tests have made the assessment of CT problematic and contentious (Ennis, 1993; Pascarella & Terenzini, 2005). Overall, much of the CT literature associates CT with certain mental processes (such as reasoning, analyzing, predicting, etc.), which can be improved through instruction, and measured by using domain-general thinking tasks that do not require specific subject matter expertise. There is a tendency toward viewing CT separate from specific subject matter knowledge, and measuring CT skills by using content that does not involve specific subject matter expertise. For example, the above-mentioned CCTT and HCTA tests do not aim to measure students' ability to think critically on specific subject matter domains, rather they use content from a variety of daily life situations that test takers at a college level are assumed to already be familiar.

One difficulty with such conceptualization of CT is that the link between performance on a domain-general CT test and specific subject matter knowledge is not clear. Some scholars (e.g. Bailin, 2002; Facione, 1990a; McPeck, 1990) argue that domain-general CT tests may measure the extent to which students are carrying out a set of mental procedures, but they are not sufficient to ensure CT proficiency. For example, when we look at Halpern's definition of CT referred to above, she explicitly defines CT as the use of thinking strategies that increase the probability of a desirable outcome. This implies that someone's use of a set of thinking strategies or procedures as such is not sufficient for CT, but equally the degree to which *a desirable outcome* is produced. Therefore, it is the quality of the outcome that distinguishes critical from uncritical thinking, and not as such knowledge of the thinking strategies.

A related issue surrounding CT assessment over the last couple of decades has been the notion of domain generality and domain specificity. The focus of the debate lies in whether CT is a set of general skills that can be applied across domains or whether it is largely specific to a particular domain (Bailin, 2002; Davies, 2013; Ennis, 1989; McPeck, 1990). The *generalist* view claims a set of CT skills exists that are general and applicable across a wide variety of domains requiring critical thought (e.g. Davies, 2013; Ennis, 1989; Halpern, 1998). On the other hand, the *specifist* view argues against the notion of general CT skills on the basis that thinking is fundamentally linked to a specific domain (e.g. McPeck, 1990). McPeck (1990) particularly contends that different domains involve different facts, concepts, and principles, and thus, skillful thought-demanding performance in one domain largely depends on having adequate knowledge and understanding of the domain in question rather than knowledge of general CT

skills. The specifist position assumes that the assessment of CT should always be pursued within the context of specific subject matter domains.

While the disagreement is longstanding, there appears to be a shift toward a synthesis of the two views (e.g. Davies, 2013). It has been acknowledged that although the related content and issues differ from one domain to the next, a set of CT skills that are transferrable across a wide variety of domains exists. Besides, the ability to think critically is understood to be highly dependent on domain-specific content knowledge, and thus an in-depth knowledge and understanding of a particular domain is required for competent performance in various thinking tasks (Davies, 2013). This implies that an accurate and comprehensive assessment of CT needs to comprise both domain-specific and domain-general CT. However, despite this theoretical claim, the assessment of CT has thus far mainly focused on domain-general CT proficiency. Although there has been a strong emphasis in designing learning environments that can promote students' domain-specific CT proficiency (e.g. Adams & Wieman, 2011; Adey & Shayer, 1994), most of these instructional attempts have not been sufficiently accompanied by reliable and valid measures of domain-specific CT proficiency. In sum, CT has mainly been linked with everyday problem-solving, and there is a general lack of experience among educators and researchers when it comes to testing for CT skills in specific science domains. The aim of this paper is therefore to develop a reliable and valid test that can measure CT in a subdomain of physics. We assume that domain-specific CT is an integral part of the expertise that specific subject matter instruction aspires toward, and define it as the ability to reasonably respond to CT tasks that require domain-specific content knowledge. We define domain-general CT proficiency as the ability to reasonably respond to CT tasks that do not necessarily require domain-specific content knowledge, but rather knowledge of everyday life.

Theoretical Background

There has been a surge of interest among various stakeholders in education to embed CT within specific subject matter instruction (for reviews, see Abrami et al., 2015; AAC&U, 2005; Tiruneh et al., 2014). The main theoretical assumption underlying the integration of CT has been that it promotes the acquisition of CT skills that can be applied to reasonably perform both domain-specific and domain-general CT tasks (Kuhn, 1999; Siegel, 1988). However, a major limitation of the existing CT assessment practice is that it emphasizes mainly on domain-general CT skills. Several researchers (see Abrami et al., 2015) examined the effectiveness of CT-embedded subject matter instruction mostly by using domain-general CT tests. It is not clearly understood from the existing CT literature whether performance in a domain-general CT test relates to mastery of a specific subject matter domain. For example, if CT skills were embedded in specific science subject and only a domain-general CT test was administered, it is not obvious from this type of research design on whether high performance in a domain-general CT test implies mastery of the specific subject matter domain in question. Similarly, administration of only domain-specific CT tests cannot give sufficient evidence on the acquisition of CT skills that can also transcend across domains. In view of a more in-depth understanding of the relationship between the acquisition of domain-specific and domain-general CT skills, it is argued in this paper that a broader

view of CT assessment is needed. Accordingly, a CT assessment framework that provides rationale for the need to assess domain-specific CT skills is described. See Fig. 1 for an overview of the proposed assessment framework. A brief description of the framework and the type of research questions that can be addressed is offered below.

As depicted in Fig. 1, the left-hand side represents the various specific subject domains or courses incorporated in a particular discipline such as physics. The assumption is that each subject matter instruction inherently targets the development of domain-specific CT skills. Therefore, for each subject domain, a corresponding CT test that measures students' ability to think critically on tasks specific to that domain needs to be administered (e.g. domain-specific CT $test_1$ for subject $domain_1$, domain-specific CT $test_2$ for subject $domain_2$, etc.). Through administration of such domain-specific tests, it is possible to measure gains in CT skills as part of students' mastery of the subject domain in question. Particularly, a research question that involves an assessment of the effect of a subject domain instruction on the development of domain-specific CT skills can be addressed with such domain-specific CT tests. It is shown in the middle section of the figure that a bidirectional link exists between each of the domain-specific CT tests. Those links depict that CT skills developed within one domain (e.g. subject $domain_1$) can transfer to a different domain (e.g. subject $domain_2$), and vice versa. Such transfer type is referred to as "near transfer" demonstrating the resemblance manifested in two separate subject domains included within a broader domain such as physics (e.g. Perkins & Salomon, 1988). Therefore, administering domain-specific CT tests may address research questions involving students' acquisition of domain-specific CT skills within specific subject matter instruction, and the extent to which acquired CT skills can transfer across subject domains. Such empirical evidence may enhance our understanding of how CT performance in one subject domain relates to CT performance in another subject domain.

The right-hand side of the figure shows a domain-general CT test that measures students' proficiency on thinking tasks that do not require specific subject matter expertise, but knowledge of everyday life. As shown in the figure, each domain-specific CT test is linked with a domain-general CT test in reverse directions. First, we assume that CT skills acquired within a specific domain may transfer to solve thinking tasks that involve knowledge of everyday life. Second, we assume that the

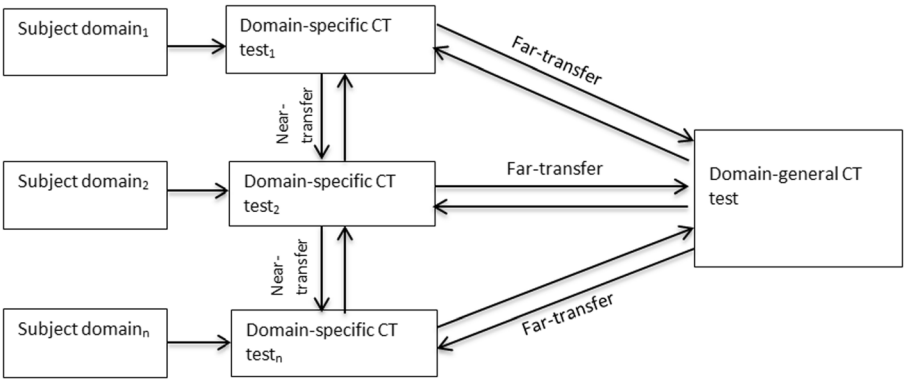


Fig. 1 A proposed framework for the assessment of domain-specific and domain-general CT skills

direct teaching of CT skills as a separate course may facilitate the acquisition of general CT skills, and those acquired domain-general CT skills may be applicable to solve thinking tasks across specific domains. Such transfer is referred to as "far transfer" because thinking tasks in a domain-general CT test may significantly differ from those in specific subject domains in terms of both surface and structural features (e.g. Perkins & Salomon, 1988).

We argue in this study that both domain-specific and domain-general CT tests need to be administered for an accurate and comprehensive understanding of the development of CT, and a better empirical understanding of the relationship that exists between domain-specific and domain-general CT proficiency. Moreover, in order to accurately examine the transferability of CT skills acquired in a specific subject domain (e.g. subject domain$_1$) to everyday life problems, both domain-specific and domain-general CT tests that target similar CT skills need to be administered. If a domain-general CT test targets, for example, problem-solving and reasoning skills in the context of real-life situations, a parallel domain-specific CT test that targets the same CT skills within the context of a specific subject domain needs to be administered. Such practice to develop domain-specific and domain-general CT tests would make it possible for researchers and educators to test for near and far transfer of CT skills.

Measuring CT in Physics: Electricity and Magnetism

A couple of CT tests exist in the broad domain of science. The Lawson's Classroom Test of Scientific Reasoning (CTSR) is the most commonly administered test in the domain of science focused on measuring scientific reasoning skills (Lawson, 1978, 2004). It is a multiple-choice test that measures scientific reasoning skills that include probabilistic reasoning, combinatorial reasoning, proportional reasoning, and controlling of variables in the context of the broader science domain (Lawson, 1978). Respondents do not necessarily need to have expertise in a specific science domain, rather the test focuses on general science-related issues that students can reasonably be presumed to have acquired in specific science subjects. The test mainly targets junior and senior high school students, but it has also been used to assess scientific reasoning skills among college science freshmen (Lawson, 1978). The other domain-specific CT test is the biology critical thinking exam (McMurray, 1991). It is a multiple-choice test with 52 items that aims to measure college students' CT skills in biology, and the items were selected from a readily available item pool developed for instructional purposes in biology.

Overall, CT has mainly been linked with everyday problem-solving, and there is a general lack of experience among researchers and educators when it comes to testing for domain-specific CT skills. To the best of our knowledge, there are no available CT tests in the domain of physics that build on students' mastery of physics. The Lawson's CTSR does measure a range of scientific reasoning skills in the context of science domains and that are based on Piaget's stages of intellectual development (Lawson, 1978, 2004). However, the CTSR focuses on the assessment of general scientific reasoning in the broad domain of science rather than the assessment of students' mastery of a specific domain of physics. The aim of this study was therefore to develop and validate a test that measures the acquisition of CT skills in physics. However, first, physics is a broad domain that deals with different subdomains at university level:

electricity and magnetism, mechanics, statistical physics, quantum mechanics, etc. It was impractical therefore to target the content from all the (physics) subdomains within one domain-specific CT test as this might result in a large amount of items, which makes it difficult to administer in a reasonable time. Second, a comprehensive CT test in physics can be administered mainly as a form of summative assessment after a student has completed some years of study in physics. Such a test cannot be used to evaluate formatively the development of students' CT skills as a result of instruction in distinct subdomains. To develop a domain-specific CT test, we therefore focused on content from a single subdomain in introductory physics, namely electricity and magnetism (E&M). We purposely selected E&M as it is a fundamental course for physics and other science majors, and an introductory course was selected as we focus on student CT development starting from the first-year of enrollment in university. In the next section, the procedures involved in developing and validating the CT test in E&M (CTEM) is presented.

Method

Planning and Development of the CTEM Test

Defining the Construct and Formulating Objectives. The first stage in developing the CTEM test was defining CT and selecting the CT skills that should be targeted in our test. As indicated in the introduction, one of the challenges in assessing CT has been the widespread disagreement among researchers and educators over what CT actually represents. We therefore initially carried out a review of the available standardized domain-general CT tests with two goals in mind. First, we aimed to identify the key CT skills that are common across various CT tests in order to have an idea of the CT skills that our test could focus on and thus establish the construct validity of the CTEM test. Second, we wanted the items for the CTEM test to mirror an existing domain-general CT test so that we can measure transfer of CT skills acquired in one domain to the other. The reviewed domain-general CT tests, as shown in Table 1, include the Cornell Critical Thinking Test–Level Z (CCTT; Ennis et al., 1985), the California Critical Thinking Skills Test (CCTST; Facione, 1990b), the Ennis-Weir Critical Thinking Essay test (Ennis & Wier, 1985), the Watson-Glaser Critical Thinking Appraisal (Watson & Glaser, 2002), and the Halpern Critical Thinking Assessment (HCTA; Halpern, 2010). These five tests were reviewed based on the following criteria: (a) is the test based on a clear definition/conception of CT, (b) are the targeted CT skills common across tests, and (c) do the test items appear to sufficiently measure the CT skills targeted on a test?

The CT skills targeted in the HCTA were selected for the CTEM test after reviewing all the above-mentioned tests in relation to the criteria by two of the co-authors. Compared to the other domain-general CT tests, the HCTA is based on CT skills that are commonly mentioned in various definitions of CT, and it includes adequate and well-structured items that appear to measure each of the identified CT skills. In addition, the HCTA is the only standardized measure of domain-general CT that uses two different types of item formats: forced-choice and constructed-response formats. The test focuses on the following elements of CT skills: reasoning, argument analysis, hypothesis testing, likelihood and uncertainty analysis, and decision-making and

Table 1 Description of commonly administered domain-general CT tests

CT instrument [a]	Targeted CT components	Item format
CCTST	Analysis, evaluation, inference, deduction, induction, and overall reasoning skills	Multiple choice
CCTT–Level Z	Induction, deduction, credibility, prediction and experimental planning, fallacies, and assumption identification	Multiple choice
Ennis-Weir CT essay test	Getting the point, identifying reasons and assumptions, stating one's point of view, offering good reasons, seeing other possibilities, and responding appropriately to and/or avoiding argument weaknesses	Essay (open-ended)
HCTA	Verbal reasoning, argument analysis, hypothesis testing, likelihood/uncertainty analysis, and problem-solving and decision-making	Both forced-choice and constructed-response
Watson-Glaser Critical Thinking Appraisal	Inference, recognition of assumptions, deduction, interpretation, and evaluation of arguments	Multiple choice

[a] The tests are aimed at measuring general CT skills and when these items were prepared, the authors used contents from a variety of real world situations that test takers are presumed to be already familiar with

problem-solving (Halpern, 2010). For each of the CT skills, Halpern developed five scenarios based on variety of everyday health, education, politics, and social policy issues. Each scenario is followed by questions that require respondents to provide a constructed response (constructed-response items), and to subsequently select the best option from a short list of alternatives (forced-choice items). In total, 25 constructed-response and 25 forced-choice items are included in the HCTA.

Once a decision was made regarding the CT skills to be included in the CTEM, the second stage involved analyzing the objectives and forms of the HCTA items. Particular emphasis was given to the constructed-response format items of the HCTA (HCTA-CR) as our goal was to develop constructed-response items to measure CT in E&M. The elaboration of the objectives of the HCTA-CR items was intended to maintain the important features of each item when constructing the CTEM items. Consistent with the objectives of the HCTA items, domain-specific CT outcomes students should demonstrate after completion of the E&M course were identified (see Table 2). The E&M learning outcomes were formulated in terms of how content experts and successful physics students can perform when confronted with CT tasks specific to E&M. It was not our intention to include in the CTEM test all the E&M learning outcomes students have to achieve in the course. The intended use of the test was rather to measure the extent to which domain-specific CT skills can be acquired as a result of a semester-based instruction in E&M. In accordance with previous suggestions on test development (e.g. Adams & Wieman, 2011), we focused on a limited set of relevant concepts incorporated in a typical introductory E&M course.

Item Format. Most of the available CT tests use forced-choice item formats. However, it is usually recommended that either a combination of both forced-choice and

Table 2 Elaboration of the domain-specific CT outcomes targeted in the CTEM test

CT category	Domain-specific CT outcomes
In the context of E&M, the student will be able to:	
Reasoning	- evaluate the validity of data
	- recognize errors of measurement
	- interpret the results of an experiment
	- detect ambiguity and misuse of definitions
Hypothesis testing	- interpret a relationship between variables
	- recognize the need for more information in drawing conclusions
	- identify when causal claims can and cannot be made
	- draw valid inferences from a given tabular or graphical information
	- check for adequate sample size and possible bias when a generalization is made
Argument analysis	- identify key parts of an argument
	- criticize the validity of generalizations in an experiment
	- judge the credibility of an information source
	- infer a correct statement from a given data set
	- identify relevant information missing in an argument
Likelihood and uncertainty analysis	- predict the probability of event
	- use probability judgments to make decisions
	- compute expected values in situations with known probabilities
	- understand the need for additional information in making decisions
	- identify assumptions (e.g. recognize what assumptions have to be maintained in generalizations from the results of an experiment)
Problem-solving and decision-making	- identify the best among a number of alternatives in solving problems
	- examine the relevance of procedures in solving scientific problems
	- recognize the features of a problem and adjust solution plan accordingly
	- evaluate solutions to a problem & make sound decisions on the basis of evidence

constructed-response, or constructed-response items only are used to measure CT skills (Ennis, 1993; Halpern, 2010; Norris, 1989). It is argued that forced-choice items do not directly and efficiently test for significant CT features, such as drawing warranted conclusions, analyzing arguments, and systematically solving problems (Norris, 1989). Based on these recommendations, we decided to measure the targeted domain-specific CT outcomes by using constructed-response items.

Item Construction. The construction of the items that can elicit the desired domain-specific CT outcomes evolved through various iterations. Efforts were made to mirror the structure of the HCTA-CR items in creating the CTEM items. In the first round of item construction, three items were created. Each of them was reviewed by all the co-authors in relation to the following criteria: (a) is the item appropriate to elicit the desired domain-specific CT performance, and (b) is the phrasing of the item clear, complete, and appropriate to the population of test takers? Eight additional items were

subsequently constructed, and a thorough revision was made based on the aforementioned criteria. After incorporating all the revisions on the eight items, construction of additional items continued using the same procedure. A couple of items from a published physics textbook were also examined and those appropriate were adapted and included in the test. The iteration continued and finally 19 items that were in line with the identified E&M CT outcomes and reviewed by all the co-authors as relevant to elicit the desired outcomes were kept. Seventeen of the items were constructed-response format items that require the test takers to demonstrate their domain-specific CT proficiency through written response, whereas two of the items were forced-choice, which require them to select from given alternatives.

Creating Scoring Guide. Parallel to item construction, the scoring guide for each item was created and reviewed by the co-authors. The scoring guide of the HCTA served as a starting point to prepare our own for the CTEM. Consistent with the objectives of each CTEM item and the types of responses expected, an ideal complete answer was initially drafted. A series of response features that could help scorers determine the extent to which specific elements are present in the student response and corresponding scoring weights were subsequently created. Item weights vary depending on the time required to fully respond to an item. Items that were agreed by the team as complex and might take more time to respond received greater weight.

Expert Review. Two physics professors and one doctoral student in the department of physics at a Flemish university were requested to review the 19 items. The three content experts were not involved in the item development. The main purpose of the CTEM test was initially explained to them, and subsequently they were requested to review each item in relation to the overall purpose of the test. Specifically, the content experts were requested to review each item based on the following criteria: (a) appropriateness of the items to the purpose of the test and the population of test takers, (b) accuracy of the information presented in the items, and (c) clarity of the words/phrases/diagrams of each item. The reviewers reported that most of the CTEM items were appropriate and relevant to measure the targeted CT skills in E&M. They had also given useful feedback on a few of the items that they thought required revision. In line with the comments, all the necessary revisions were made.

Student Cognitive Interviews and Small-Scale Paper-Pencil Administration. Cognitive interviewing is a method used to examine whether respondents comprehend and respond to items the way researchers intend (e.g. Willis, 2005). After incorporating all the expert comments, audiotaped cognitive interviews were conducted with a small group of second-year physics major students ($N = 6$). We sent emails to six students, who had recently completed an E&M course and agreed to participate in the interviews. The goal of the cognitive interview was to examine whether the students would respond to the CTEM items as intended. Prior to the interviews, we prepared an interview protocol that included requesting each interviewee to (a) briefly introduce him/herself (e.g. name, year of study, when exactly he/she was enrolled in the E&M course), (b) read aloud an item and immediately mention if there were words/phrases/drawings that are difficult to understand, (c) think out loud while giving a response to an item, and (d) give an overall estimation of whether the items included in the test were easy or difficult to solve, and

why. One interviewer, assisted by one of the co-authors, conducted the interviews. The interviewer started by giving a brief explanation regarding the purpose of the interview, and requested an interviewee to introduce him/herself. Each participant was then asked to read aloud an item and state if there had been difficulty to understand the overall ideas of the item (e.g. underline or circle words/phrases difficult to understand). The first two phases of the interview protocol served to set the stage for the think aloud phase. Each of the interviewee was subsequently asked to solve each item through thinking out loud. Some probing questions were used by the interviewer to elicit the required response. As suggested by Adams and Wieman (2011), efforts were made to put the interviewees in "authentic test-taking mode" (p. 14), with a great caution not to intervene with students' explanation. After completion of the test, each interviewee was requested to give an overall reflection on the entire test and/or specific items. For instance, two interviewees skipped a few questions during the problem-solving phase. The interviewer asked explanations on why they skipped the items.

The interviews were later transcribed and analyzed. The findings were categorized along two dimensions: level of understanding (vocabulary, sentence structure, length of statements, clarity of drawings and instructions, interpretation), and overall accuracy of students' response. The findings revealed that 10 of the 19 CTEM items were clear and easily understood by all the interviewees. However, significant revisions were made on the other 9 items that involved the following: (a) rephrase words and phrases that were either ambiguous or difficult to understand, (b) shorten items that were too long, (c) include further explanatory instructions, and (d) improve the clarity of the figures and tables that were misinterpreted.

Parallel to the cognitive interview, preliminary version of the CTEM test was pilot tested to a small group of second-year physics major students ($N = 19$) in a Flemish university. The primary purposes of the pilot testing were to examine whether the items could elicit the desired responses in an authentic testing setting, determine whether test responses could easily be scored with the proposed scoring guide, and get an initial indication of the time required to complete the test. Analysis of the students' responses indicated that some items needed significant revision. It was revealed that the students' responses to a few of the items lacked clarity. In addition, the pilot testing also helped to revise the scoring guide. During the initial design of the scoring guide, the possible answers suggested for a few items were limited. Through the pilot testing, however, a wide range of responses were discovered for a few items that appeared to be relevant. Accordingly, revisions on the scoring guide were made. Administration of the test lasted between 50 and 60 min.

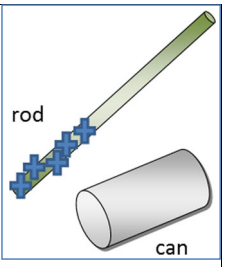
Item Revision and Administration. Revisions were made on the preliminary version of the items based on data from the cognitive interview and pilot test. The cognitive interview particularly indicated that one of the items was lengthy and difficult for the students to understand. Therefore, a decision was made during the revision to split that particular item into two separate items. The revised version of the CTEM test therefore consisted of 20 items, of which 18 are constructed-response and the remaining forced-choice format items (see Fig. 2 for sample CTEM item and corresponding scoring guide).

After incorporating all the revisions, the revised version of the test was administered to second-year students ($N = 45$) with major in electromechanical engineering in a Flemish university. These students did not participate in any of the previous pilot

Sample item: Hanna does the following experiment: she brings a positively charged rod close to a metal can. Doing the experiment shows that the can is attracted to the rod.

Hanna is puzzled with the result of her experiment. She expected the negative electrons on the metal would be attracted to the rod while the positive nuclei are repelled, and opposite forces cancel out, which would mean that the can remains at rest.

How can you make Hanna's argument consistent with the experiment? Indicate all the possible explanations.



Item objective:
The sample item measures one of the sub-skills within the argument analysis CT element: identify key parts of an argument and relevant information missing in an argument (see Table 2). The item requires that the student criticizes or evaluates the general statement that 'opposite forces cancel out' and identifies clearly the argument misses the relevant information on distance effect in Coulomb's law.

Scoring guide for the sample item [item weight: 4 points]:

The following is an ideal complete answer that we expect from a student:

> *The positively charged rod draws the loosely bound electrons and accumulates them at the side of the can closest to the rod while leaving the other side positively charged. Since the distance between the rod and the negatively charged side of the can is smaller than the distance between the rod and the positively charged side of the can, the attractive force between the rod and the can is larger than the repulsive force between them. According to Coulomb's law, the force is inversely proportional to the square of the distance. Therefore, the net force on the can is attractive.*

Scoring guide:

- Does the student answer refer to the motion of electrons?
  If yes award 1 point
- Does the student answer refer to $F_{attract} > F_{repel}$?
  If yes award 1 point
- Does the student answer refer to distance effect in $F_{coulomb}$?
  If yes award 2 points, but if the answer mentions distance irrespective of its relationship with force, award only 1 point.

Fig. 2 Sample CTEM item and corresponding scoring guide

studies and were enrolled in an E&M course about 6 months prior to the CTEM test administration. The test participants consisted of 42 men and 3 women between the ages of 19 and 23 years ($M = 20.04$, $SD = 1.17$). Prior to the test administration, the students were provided oral instruction regarding the purpose of the test, general direction on how they should respond to the items, and a request to take the test seriously. The test was administered in a controlled classroom setting and great caution was made for all students to hand in the test so that test questions would not circulate. There was no strict time restriction to complete the test, but students were told at the beginning that it might take about an hour to complete. About 80 % of the students were able to finish within 50 min and the remaining completed in 60 min.

To assess the convergent validity of the CTEM, the HCTA was administered to the same participants immediately after they completed the CTEM. Although the content for the CTEM and HCTA tests differ from each other, as noted above, both tests focus on similar CT skills. Both the original and translated version of the HCTA was validated by researchers in different countries such as the USA (Halpern, 2010), Belgium (Verburgh, François, Elen, & Janssen, 2013), and China (Ku & Ho, 2010). For the purpose of this study, we administered the original (English) version. It was not the purpose of this study to validate the HCTA, but rather to evaluate how performance on the CTEM relates to the HCTA. Cronbach's alpha coefficient was computed to measure the internal consistency of the HCTA for the present participants and it was found to be acceptable ($\alpha = .75$).

Results

In this section, we describe the results of our analysis of the CTEM test including the inter-rater reliability, internal consistency, item difficulty, item discrimination, and convergent validity.

Inter-Rater Reliability

To evaluate the inter-rater reliability, we computed the CTEM test results from 15 randomly selected test participants that were scored independently by 2 different raters using the same scoring guide. The inter-rater agreement for each of the 18 constructed-response items range from .71 to 1.00 (weighted Cohen's kappa coefficient). See Table 3 for an overview of the kappa values. For item 1, all the test takers from the randomly selected 15 responses scored "0" and therefore the kappa coefficient could not be computed. For the total scores of the test, the inter-rater agreement was .83. The results overall showed sufficient to high inter-rater reliabilities both at the item and the total score level. In addition, paired sample t test was computed to examine the effect of the rater on the mean scores of each item. A Shapiro-Wilk's test ($p > .05$) and a visual inspection of the histograms and box plots showed that the scores by the two raters were approximately normally distributed. The results of the paired sample t test indicated no statistically significant difference between the scores allocated to each item by the two raters ($p > .05$). This indicates the scoring objectivity of the test and that it is meaningful to compare students' performance on individual items as well as on the test entire test itself. Table 4 shows an overview of sample student responses and awarded scores for item 5 of the CTEM test.

Internal Consistency

Cronbach's alpha coefficient was computed to measure the internal consistency of the CTEM test and it was found to be acceptable, $\alpha = .72$. Although a desirable value for internal consistency may vary as a function of the nature of the construct being measured, Cronbach's alpha values between .7 and .8 are considered acceptable (Cohen, Manion, & Morrison, 2007; Nunnally, 1978).

Table 3 Summary statistics for the CTEM items

Item	Cohen's kappa coefficient ($n = 15$)	Item difficulty	Item discrimination
Item 1[a]	–	.17	.16
Item 2	.89	.51	.35
Item 3	.86	.36	.28
Item 4	.81	.27	.36
Item 5	.87	.25	.23
Item 6	.77	.29	.27
Item 7[b]	–	.61	.25
Item 8[b]	–	.49	.31
Item 9	.91	.47	.50
Item 10	.77	.30	.25
Item 11	.71	.22	.23
Item 12	.87	.40	.60
Item 13	.87	.27	.18
Item 14	.74	.28	.30
Item 15	1.00	.21	.17
Item 16	.76	.35	.23
Item 17	.80	.38	.33
Item 18	.78	.22	.18
Item 19	.79	.33	.22
Item 20	.89	.31	.24

[a] Among the randomly selected 15 test responses, all students scored "0" for this item and kappa cannot be computed

[b] Forced-choice format items

Item Difficulty and Discrimination

To establish an additional measure of the features of the CTEM test, we computed the difficulty and discrimination power of each item. The possible scores for each of the constructed-response CTEM items range from 0 to 10. For instance, the possible score for item 1 ranges from 0 to 2; for item 18 from 0 to 10; and for item 20 from 0 to 5. To compute the item difficulty and item discrimination, we used the formula suggested for open-ended items by the Evaluation and Examination Service of the University of Iowa (EES, 2004).

Below is the formula we used to compute the item difficulty (P):

$$P = \frac{\sim fX - nX_{min}}{n(X_{max} - X_{min})}$$

where ~fX is the total number of points earned by all students on an item, n is the number of students, $X_{\min}$ is the smallest item score possible, and $X_{\max}$ is the highest item score possible.

Table 4 Sample student responses and awarded scores

Sample student responses for the CTEM item in Fig. 2 and the scores awarded are illustrated below:	
Student 1:	The electrons move through the two bodies. If we approach the can with the rod, more electrons will move to the surface causing the rod and the can to attract each other. We get a redistribution of charges. As the rod is positively charged, there are more protons in the rod than electrons.
	→Awarded 1 point (motion of electrons = 1)
Student 2:	Because negative electrons are attracted and positive ions are repelled, the can will have a positive and a negative side. As the electric force decreases with distance ($F \sim 1/r^2$), the negative side will be attracted more strongly that the positive side is repelled. The can will move toward the rod because there is net force toward the rod.
	→ Awarded 4 points (motion of electrons = 1, $F_{attract} > F_{repel} = 1$, $F_{coulomb} = 2$)
Student 3:	The amount of positive kernels and negative electrons are not equal, the rod has much more positive kernels thus the net force is not zero.
	→ Awarded 0 points
Student 4:	The electrons in the can move toward the rod, so the average distance between the rod and the electrons is smaller than the average distance between the rod and positive particles. There is a non-zero resultant force.
	→ Awarded 2 points (motion of electrons = 1, distance mentioned = 1)

We computed the difficulty level by incorporating all the test participants ($N = 45$). As shown in Table 3, the difficulty indices for the CTEM items range from .17 to .61. Although the ideal item difficulty depends on the purpose of the assessment (Educational Testing Service [ETS], 2008), it appears that most of the CTEM items were rather difficult to our test participants. About 50 % of the items were found to have difficulty index of less than .30. As indicated in Table 2, we formulated CT outcomes that could be focused on specific subject matter domains. The difficulty level was computed to get an idea of the proportion of test takers correctly responding to an item. However, it is important to recognize that the CTEM test is more of a criterion-referenced test that aims to measure acquisition of domain-specific CT skills. For such criterion-referenced assessments, a test may be exceedingly difficult for particular test takers and still be appropriate (ETS, 2008). Therefore, the item difficulty indices at this initial stage of test validation may not necessarily lead to revision or exclusion of the items or scoring criteria.

To compute the item discrimination index, the item scores of groups of high and low scoring students were selected. As item discrimination describes how well an item can distinguish between individuals with different levels of ability, the upper (U) and lower (L) groups were selected from the extremes of the score distribution. Although the most common approach to U-L grouping is to take the highest and lowest 27 % of the test scores, we had limited number of participants ($n = 45$) and decided to use the U and L 22 % of the scores to compute the item discrimination indices.

The discrimination index is calculated using the following formula (EES, 2004):

$$\text{Item discrimination}(D) = P_U - P_L$$

where P_U and P_L are the difficulty indices for the U and L groups.

As shown in Table 3, the item discrimination indices range from .16 to .60. Although we recognize that the optimum discrimination index varies with the purpose of the test and type of item format (ETS, 2008), four of the CTEM items were found to have discrimination index of less than .20. A possible reason for such very low discrimination index values may relate to the corresponding item difficulty index of each item. It is revealed that those particular items with lower discrimination indices were relatively more difficult than the rest of the items (see Table 3). There is evidence that items that are very easy or very difficult will appear less discriminating (Schmidt & Embretson, 2003).

In order to explore how the subscores function within an item, an additional upper-lower group analysis at the level of the subscores was performed. Taking the L and U 22 % of the score distributions, the frequencies of the subscores for each item were plotted (see Fig. 3a, b for sample illustrations to item 1 and item 2, respectively). The maximum score for each of the two items is 2, and as can be seen in the figures, the higher subscores increase from the lower to the upper score group, whereas the lower subscores decrease. It is the same for all the remaining 18 items, even for the 4 items with lower discrimination indices. This additional score group analysis further supports the discriminatory value of the CTEM items.

Convergent Validity

Evidence for the convergent validity of an assessment instrument can be derived by showing the correlation of the assessment instrument with an existing standardized test. The participants' CTEM performance was compared therefore to their performance on the HCTA test. It was predicted that these two tests would result in a significant positive correlation between the two sets of scores as both of them focus on similar set of CT skills. A visual inspection of the histograms, box plots, and a Shapiro-Wilk's test ($p > .05$) of the scores on the two tests showed no violation of linearity and normality. Calculation of the Pearson's correlation coefficient showed a significant positive correlation between the two sets of scores ($r = .45$, $p = .02$, $N = 45$). In addition, an estimate of the correlation at the true score level was obtained by calculating the correlation after correction for attenuation (Osborne, 2003).

$$r_{x'y'} = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}}$$

The two variables that we were interested to measure, the CTEM and the HCTA, are represented by x and y, respectively. In the above equation, r_{xy} is the observed

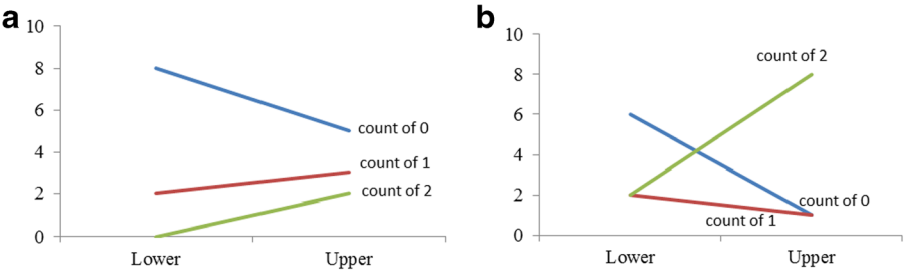


Fig. 3 **a** Upper-lower group score distributions for item 1. **b** Upper-lower group score distributions for item 2

correlation between the CTEM and HCTA, r_{xx} and r_{yy} refer to the reliability coefficients of the CTEM and HCTA respectively, for which the value of Cronbach's α was used as a lower bound to estimate. It was found that the correlation between the CTEM and HCTA after corrected for attenuation is .62. The relative increase in the size of the correlation coefficient implies that the two tests are measuring the same construct, and thus suggests the CTEM has adequate convergent validity (e.g. Jöreskog, 1971).

Discussion and Conclusions

As the importance of developing students' ability to think critically on specific domains of science continues to grow, researchers and practitioners need to have valid and reliable tests to evaluate the effectiveness of various instructional efforts. In this study, we argued that an accurate and comprehensive assessment of CT should emphasize both domain-specific and domain-general CT dimensions. Recognizing the lack of domain-specific CT tests in the science domain, a test that can assess useful elements of CT in E&M was systematically developed and validated. Analyses of the qualitative and quantitative data overall provide sufficient evidence that the CTEM test at this initial stage can be a good basis for measuring CT skills in E&M. A systematic review of some of the available domain-general CT tests was conducted to identify the core CT skills that are common across various CT tests, and thus the CTEM items were designed to mirror the five CT skills included in the HCTA. Content experts were involved during the item development stage in reviewing the items, and cognitive interviews were conducted with selected students, which provided evidence that the test items were clear and elicited the targeted domain-specific CT outcomes.

Moreover, the quantitative evidence showed that the CTEM test produced a sufficient inter-rater agreement and acceptable reliability coefficient. However, it has to be noted that the coefficient alpha was not as large as expected. The relatively low alpha value can be explained by at least two factors. First, the CTEM items were intended to elicit students' ability to demonstrate the five targeted CT outcomes as outlined by Halpern (2010): reasoning, argument analysis, hypothesis testing, likelihood and uncertainty analysis, and problem-solving and decision-making. It is possible that the cognitive processes required to respond to the items were multifaceted. For instance, a student who performed well in an item that focuses on argument analysis may not have done well on a different item that focuses on hypothesis testing as these two components slightly vary in terms of the required cognitive processes. Second, the lower coefficient alpha may have to do with the composition of the test group and number of participants. As noted above, the CTEM test was administered to a group of students with a major in mechanical engineering and who were enrolled in an E&M course 6 months prior to the test administration. Responding accurately and consistently to the CTEM items require students to have an adequate mastery of the E&M content. The present test takers clearly found most of the items very difficult and may have possibly responded to some of the items randomly. This may have influenced the internal consistency of the test. The answers to the CTEM items is expected to become more consistent and refined when the test is given immediately following completion of a well-designed E&M instruction.

Although the test was relatively difficult, it showed sufficient discriminatory value, as evidenced by the discrimination indices and the additional score group analysis. Given the complexity of the construct we aim to measure, we do not suggest at this stage to exclude any

of the CTEM items. As all the CTEM items were evaluated very useful in measuring the targeted CT outcomes during the expert review and student cognitive interviews, excluding a couple of them at this stage could risk the content validity of the test. Additional validation studies that involve a larger and diverse group of respondents representing the target population should be conducted to further strengthen the quantitative data set and related measures.

The procedures described in this study to develop and validate the CTEM items are largely in line with the guidelines suggested for the preparation of constructed-response and other performance tests (e.g. Adams & Wieman, 2011; Aydın & Ubuz, 2014; Benjamin et al., 2015). Although the item development and validation procedures were based on established guidelines of previous research, this study has proposed a CT assessment framework that may promote the measurement of both domain-specific and domain-general CT skills. Our hope is that the CTEM test can be used both for instructional and research purposes. First, it can be used to address research questions that involve an assessment of the effectiveness of instructional interventions on the acquisition of domain-specific CT skills. Through this test, researchers and educators can examine the extent to which an instructional intervention stimulates the acquisition of domain-specific CT skills in E&M over the course of a semester. We recognize that examining the effect of systematically designed instructional interventions for the acquisition of domain-specific CT skills may require administering domain-specific CT tests, for instance, the CTEM both before and after an intervention. However, given the requirement that sufficient content knowledge is needed to adequately respond to a domain-specific CT test, administering such tests prior to an instructional intervention is impractical. For instance, it is impossible to administer the CTEM test prior to instruction in an introductory E&M course because students may not have yet acquired adequate mastery of the subject matter content. One possible recommendation for researchers who wish to measure domain-specific CT skills prior to subject domain instruction is to administer a parallel test with the same CT skills but by using content that test takers are assumed to already be familiar. Second, the CTEM can be used to address research questions that focus on the relationship between the acquisition of domain-specific and domain-general CT skills. Since the CTEM items are designed to mirror an already established and standardized domain-general CT test (viz., the HCTA), it is possible to compare students' performance on the CTEM and the HCTA. To the best of our knowledge, such systematic attempts of mapping a domain-specific CT test to a standardized domain-general CT test have not been published yet.

Although we acknowledge that the validation procedures described in this study represent a first attempt, we hope to have demonstrated an approach that can be applied to develop and validate CT tests in other domains in the science and arts. Such practice of mapping domain-specific and domain-general CT tests may reinforce the notion that both domain-specific and domain-general CT skills need to be jointly targeted in the assessment of CT.

Acknowledgments We thank An Verburgh for her assistance in collecting data at the initial phase of the test validation. We also thank the physics experts who provided us feedback on the first version of the CTEM test, and Jeroen Buijs for his support in administering the revised version of the test.We would also like to express our deepest appreciation to the anonymous reviewers for the constructive comments and suggestions on an earlier version of the manuscript.

References

Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, A. & Persson, T. (2015). Strategies for teaching students to think critically: a meta-analysis. *Review of Educational Research, 85*(2), 275–314. doi:10.3102/0034654314551063.

Adams, W. K. & Wieman, C. E. (2011). Development and validation of instruments to measure learning of expert like thinking. *International Journal of Science Education, 33*(9), 1289–1312. doi:10.1080/09500693.2010.512369.

Adey, P. & Shayer, M. (1994). *Really raising standards: Cognitive intervention and academic achievement*. London, UK : Routledge.

Association of American Colleges and Universities. (2005). *Liberal education outcomes: A preliminary report on student achievement in college. Liberal education*. Washington, DC: AAC&U.

Aydın, U. & Ubuz, B. (2014). The thinking-about-derivative test for undergraduate students: development and validation. *International Journal of Science and Mathematics Education, 13*(6), 1279–1303. doi:10.1007/s10763-014-9545-x.

Bailin, S. (2002). Critical thinking and science education. *Science & Education, 11*, 361–375.

Benjamin, T. E., Marks, B., Demetrikopoulos, M. K., Rose, J., Pollard, E., Thomas, A. & Muldrow, L. L. (2015). Development and validation of scientific literacy scale for college preparedness in STEM with freshman from diverse institutions. *International Journal of Science and Mathematics Education*. Advanced online publication. doi:10.1007/s10763-015-9710-x.

Cohen, L., Manion, L. & Morrison, K. (2007). *Research methods in education* (6th ed.). London, UK: Routledge.

Davies, M. (2013). Critical thinking and the disciplines reconsidered. *Higher Education Research and Development, 32*(4), 529–544. doi:10.1080/07294360.2012.697878.

EES (2004). *Preparing and evaluating essay test questions: Technical bulletin #36*. Retrieved October 23, 2014 from http://www.uiowa.edu/~examserv/resources_fees/Technical_Bulletins/TechBulletin 36.pdf.

Ennis, R. H. (1989). Critical thinking and subject specificity: clarification and needed research. *Educational Researcher, 18*(3), 4–10. doi:10.3102/0013189X018003004.

Ennis, R. H. (1993). Critical thinking assessment. *Theory Into Practice, 32*, 179–186.

Ennis, R. H., Millman, J. & Tomko, T. N. (1985). *Cornel critical thinking test level Z*. Pacific Grove, CA: Midwest publications.

Ennis, R. H. & Wier, E. (1985). *The Ennis-Wier critical thinking essay test*. Pacific Grove, CA: Midwest publications.

ETS (2008). *Guidelines for constructed-response and other performance assessments*. Retrieved October 21, 2014 from http://www.ets.org/Media/About_ETS/pdf/8561_ConstructedResponse_guidelines.pdf.

Facione, P. A. (1990a). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. Research findings and recommendations.* Retrieved from ERIC database. (ED315 423)

Facione, P. A. (1990b). *The California critical thinking skills test - college level. Technical report #2. Factors predictive of CT skills*. Millbrae, CA: California Academic Press.

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains. *American Psychologist, 53*(4), 449–455. doi:10.1037//0003-066X.53.4.449.

Halpern, D. F. (2010). *The halpern critical thinking assessment: Manual*. Modling, Austria: Schuhfried GmbH.

Halpern, D. F. (2014). *Thought and knowledge: An introduction to critical thinking* (5th ed.). New York, NY: Psychology Press.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*(2), 109–133. doi:10.1007/BF02291393.

Ku, K. Y. L. & Ho, I. T. (2010). Dispositional factors predicting Chinese students' critical thinking performance. *Personality and Individual Differences, 48*(1), 54–58. doi:10.1016/j.paid.2009.08.015.

Kuhn, D. (1999). A developmental model of critical thinking. *Educational Researcher, 28*(2), 16. doi:10.2307/1177186.

Lawson, A. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching, 15*(1), 11–24. doi:10.1002/tea.3660150103.

Lawson, A. (2004). The nature and development of scientific reasoning: a synthetic view. *International Journal of Science and Mathematics Education, 2*(3), 307–338. doi:10.1007/s10763-004-3224-2.

Lin, S.-S. (2014). Science and non-science undergraduate students' critical thinking and argumentation performance in reading a science news report. *International Journal of Science and Mathematics Education, 12*(5), 1023–1046. doi:10.1007/s10763-013-9451-7.

McMurray, M. A. (1991). Reliability and construct validity of a measure of critical thinking skills in biology. *Journal of Research in Science Teaching, 28*(2), 183–192.

McPeck, J. (1990). *Teaching critical thinking: Dialogue & dialectic*. New York, NY: Routledge.

Niu, L., Behar-Horenstein, L. S. & Garvan, C. W. (2013). Do instructional interventions influence college students' critical thinking skills? A meta-analysis. *Educational Research Review, 9*, 114–128. doi:10.1016/j.edurev.2012.12.002.

Norris, S. P. (1989). Can we test validly for critical thinking? *Educational Researcher, 18*(9), 21–26. doi:10.2307/1176715.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Osborne, J. W. (2003). Effect sizes and the disattenuation of correlation and regression coefficients: lessons from educational psychology. *Practical Assessment, Research & Evaluation, 8*(11), 1–11.

Pascarella, E. T. & Terenzini, P. T. (2005). *How college affects students: A third decade of research* (Vol. 2). San Francisco, CA: Jossey-Bass.

Perkins, D. & Salomon, G. (1988). Teaching for transfer. *Educational Leadership, 46*(1), 22–32. Retrieved from http://arxiv.org/abs/0704.1854.

Schmidt, K. M. & Embretson, S. E. (2003). Item response theory and measuring abilities. In J. A. Schinka & I. B. Weiner (Eds.), *Handbook of psychology: Research methods in psychology* (Vol. 2, pp. 429–444). Hoboken, NJ: Wiley.

Siegel, H. (1988). *Educating reason: Rationality, critical thinking, and education*. New York, NY: Routledge.

Tiruneh, D. T., Verburgh, A. & Elen, J. (2014). Effectiveness of critical thinking instruction in higher education: a systematic review of intervention studies. *Higher Education Studies, 4*(1), 1–17. doi:10.5539/hes.v4n1p1.

Verburgh, A., François, S., Elen, J. & Janssen, R. (2013). The assessment of critical thinking critically assessed in higher education: a validation study of the CCTT and the HCTA. *Education Research International, 2013*(1), 1–13. doi:10.1155/2013/198920.

Watson, G. & Glaser, E. (2002). *Watson-Glaser critical thinking appraisal*. London, UK: Pearson Assessment.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.